Dynamic Segmentation of Vocal Extract for Assamese Speech to Text Conversion using RNN

Krishna Dutta and Kandarpa Kumar Sarma
Department of Electronics and Communication Technology
Gauhati University
Guwahati-781014, Assam, India
e-mail: krishnadutta54@gmail.com, kandarpaks@gmail.com

Abstract — The current work proposes a prototype Speech to Text Conversion System (STSC) in Assamese language using Linear Predictive Coding (LPC) and Recurrent Neural Network(RNN). The LPC features are extracted from utterances of isolated phonemes of Assamese language (a major language of North-East India). These are used to train a RNN by a proposed dynamic method. The proposed method segments an utterance with an optimal dynamic criterion to improve the success scores during testing of the STCS system. The proposed method dynamically adjusts the length of the windows required for recognizing different phonemes. The performance of the proposed method is compared with a conventional static RNN based STCS system which is trained using prior knowledge about length of windows required for recognizing different phonemes.

Index Terms—SCTS, Dynamic Segmentation, Moving Average Filter, LPC, RNN.

I. Introduction

Recognition of speech input is an essential component in a Speech to Text Conversion System (STCS). Recognition performance, however, suffers if nearby speech samples are not successfully separated out. The separation of speech contents in a vocal capture plays a critical part in any speech processing application. The precision in this segmentation subsequently determines the performance of a STCS. In most available STCS, such segmentation is performed statically. Some of the reported works are [1] [2] . In this approaches, the STCS suffer when input speech segments are continuously related [3]. We here propose a method which performs extraction of voiced and unvoiced speech segments (or phonemes) dynamically. The proposed system is a dynamic segmentation mechanism designed using a Recurrent Neural Network (RNN) which separates speech components by adaptively fixing section boundaries. The system is tested under a range of conditions from which satisfactory performance is achieved under different test conditions. The paper is organized as follows: Section II describe the phonological features of Assamese language. The proposed dynamic speech segmentation technique is describe in Section III. In Section IV experimental details are summarized. Finally, conclusion and further directions are included in Section V.

II. Distinctive Phonological Features of Assamese Language

Assamese is a major speaking language in North-East India particularly in Assam with its own unique identity, language and culture. Its origins root back to the Indo-European family of languages. These languages are spoken by more than a billion people, chiefly in Afghanistan, Bangladesh, India, Iran, Nepal, Pakistan, and Sri Lanka. It also is related to the Indo-Iranian subfamily. This class can be subdivided into three groups of languages: the Dardic(or *Pisacha*), the Indic (or Indo-Aryan), and the Iranian. Assamese is the easternmost member of this New Indo-Aryan (NIA) subfamily spoken in the Brahmaputra Valley of Assam [4]. Retaining certain features of its parent Indo-European family it has got many unique phonological characteristics. Some of those may be cited as below:

- A unique feature of the Assamese language is a total absence of any retroflex sounds. Instead the language has a whole series of alveolar sounds, which include oral and nasal stops, fricatives, laterals, approximants, flaps and trills, unlike other Indo-Aryan and Dravidian languages [5].
- Another striking phonological feature of the Assamese language is the extensive use of velar nasal / η/. In other New Indo Aryan languages this / η/ is always attached to a homorganic sound like /g/. In contrast it is always used singly in Assamese.
- The voiceless velar fricative / x / is a distinct characteristic of Assamese language which is not to be found in any language in the entire country. It is similar to the velar sound in German of Europe. Phonetically, this /x/ sound is pronounced somewhat in between the sounds /s/, /kh/ and /h/ and is similar to the German sound /ch/ as pronounced in the word Bach or the Scottish sound as found in the word Loch. It may be an Indo- European feature, which has been preserved by 'Asomiya'. It is an important phoneme in the language [3] [4] [5] .

There are other phonological uniqueness of Assamese pronunciation which shows minor variations when spoken by people of different regions of the state. This makes Assamese speech unique and hence requires a study exclusively directly to develop a speech recognition / synthesis system in Assamese.

978-1-4577-0720-9/12/$26.00 ©2012 IEEE

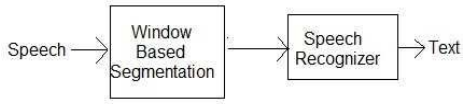

Fig. 1. Conventional Speech to Text Converter

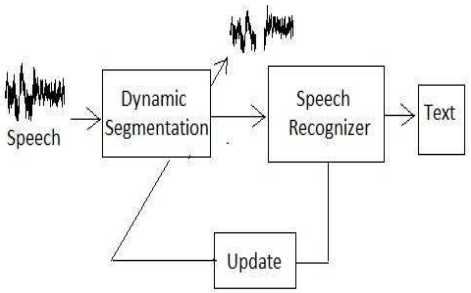

Fig. 2. Dynamic Segmented based Speech to Text Converter

III. Dynamic Speech Segmentation- Proposed System Model and Related Considerations

A generic STCS is shown in Fig 1. Prior information of input sample is the prime requirement for high success rate in this conventional method. Using this prior information, the length of the window is fixed. It helps in segmenting the speech signal. This limitation is completely resolved by our proposed method where the windowing concept is replaced by a dynamic segmentation and thresholding method. In the proposed method, the requirement of prior information is eliminated. The block diagram of the proposed method is shown in the Fig 2. The steps of the algorithm are summarized below-

- Suppress the short term fluctuation and noise by using a moving average (MA) filter.
- Starting and ending position of speech segment of the mixed input sample is determined by thresholding.
- Segmented portion of the speech segment is applied to the recognizer.
- If it is unable to recognize, then small portion is added to the previous one and again sent to the recognizer.
- This process is repeated until the recognizer finds the best match.
- When it finds a match, leaves the previous portion and send a new one to the recognizer by following the previous steps.
- This process continue until the sample is finish.

The process logic of the above steps is expressed by a flowchart shown in Fig 3.

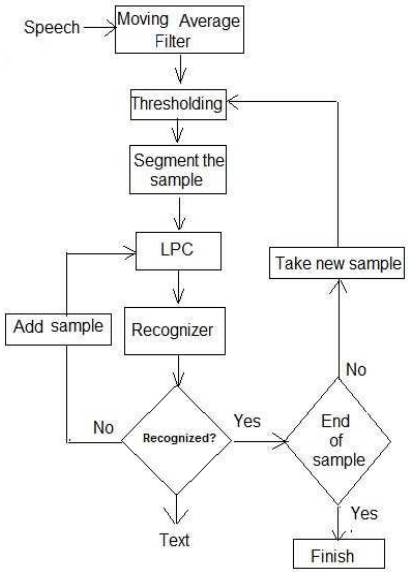

Fig. 3. Graphical view of the proposed algorithm

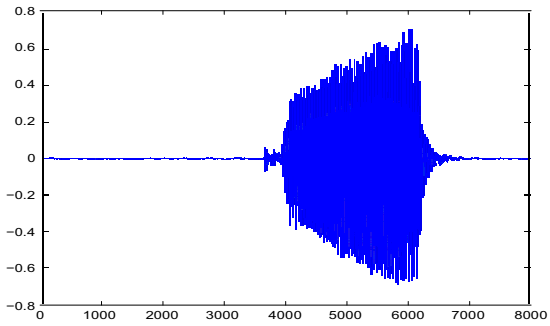

Fig. 4. Isolated Speech Signal(("*Ko*") k in English)

A. Raw speech signals

As mentioned earlier, since the work is a part of Assamese STCS, speech signals of uttering Assamese letters(vowels and consonants) are captured. In this process some isolated mixed speech signals are also captured. The letters are used to train the speech recognizer. In the meantime, concatenated isolated speech signals are used to check the flexibility of the system by segmenting and recognizing the segmented speech. The spectrum of isolated and concatenated isolated speech signals are shown in Figure 4 and Figure 5 .

For recording the speech signal, a PC headset and a sound recording software, Gold Wave, is used. GoldWave's Monitor recording option helps to adjust the sampling rate and duration of recording. While recording, the sampling rate taken is 8000 Hz in mono channel mode.

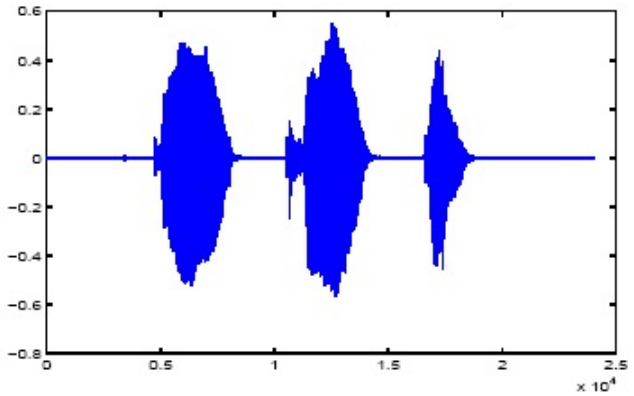

Fig. 5. Concatenated isolated speech signals

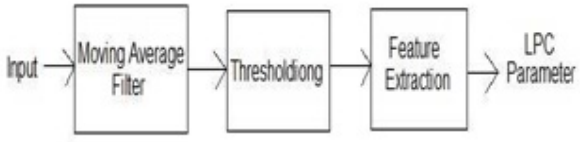

Fig. 6. Block diagram of Linear Predictive Coding

B. Moving Average Filter

Moving average filter is a type of finite impulse response filter. It analyzes a set of data points by creating a series of averages of different subsets of the full data set. Mathematically, a moving average filter is expressed as-

$$MA_n = \frac{(p_n + p_{n-1} + \ldots\ldots\ldots + p_{n-m})}{m} \qquad (1)$$

where m is the size of the subset and n is the data point for calculated moving average. A MA filter is commonly used to smooth out short-term fluctuations normally seen in the recording of speech signal. The threshold between short- term and long-term depends on the application, and the parameters of the moving average are set accordingly [7]. In this experiment, we take m is equal to 10 to suppress the short-term fluctuation.

C. Thresholding

Thresholding is a method of segmentation normally used in image processing applications. In speech processing, thresholding is also applicable. In this case, energy thresholding technique is used. By simply setting a threshold to the sample, the starting and ending position of the speech section can be identified. This leads to the replacement of the windowing method. In Fig 7, 'Speech' and 'Silence' segments are shown. The separated speech segment is shown in Figure 8. After the normalization, 0.1 dB is set as a threshold for segmentation.

D. Extraction of voice features

Feature is a set of values extracted from an input speech that uniquely represents the key attributes of the sample. Speech signal is the output of a time varying vocal tract stimulated by a time-varying excitation signal. The vocal tract system is approximately described in terms of the acoustic features such as the frequency response of the resonances (formants) and anti- resonances (anti-formants) of the systems. These features are easier to extract from the signal than the articulatory parameters. The excitation of the vocal tract consists of broadly three categories [3][6]: voiced source (due to vibrating vocal folds), unvoiced source (turbulent air flow at narrow constriction in the vocal tract) and plosive source (abrupt release).

In general, the short-time characteristics of the speech signal are represented by the short-time (10-20 mS) spectral features of the vocal tract system as well as the nature of excitation in the short-time segment. These are called segmental features [6]. Selecting appropriate features from a speech signal is an important issue in speech recognition. In this process, separation of speech component from the raw input speech is one of the primary concerns of achieving high accuracy.

E. Linear Predictive coding

Linear predictive analysis of speech has become the predominant technique for estimating the basic parameters of speech. Linear predictive analysis provides an accurate estimate of the speech parameters and also an efficient computational model of speech. The basic idea behind linear predictive analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences, over a finite interval between the actual speech samples and linear predicted values a unique set of parameters or predictor coefficients can be deter- mined [8].

IV. Results and Discussion

In this section a brief description of the experimental consideration and related results are given.

A. Size of the feature set as fixed by the LPC predictor

The order of the LP analysis used for generating the LPC feature set is an important factor. If the feature set is selected without any logic, the results vary. Experiments are carried out to optimize the length of the feature vector for generating the corpus. Table I shows the effect of predictor length in generating acceptable recognition performance by an ANN. The LPC predictor size of 35 gives the best results from the RNN during training [8]. Its successful recognition rate and the time required to reach the desired mean square error(MSE) value is the best among the eight cases considered. The time-precision combination of the 35^{th} order LPC prediction when applied to an RNN generates the best precision level in recognizing the input samples. If the LPC feature vector size is very long, the training time becomes very long.

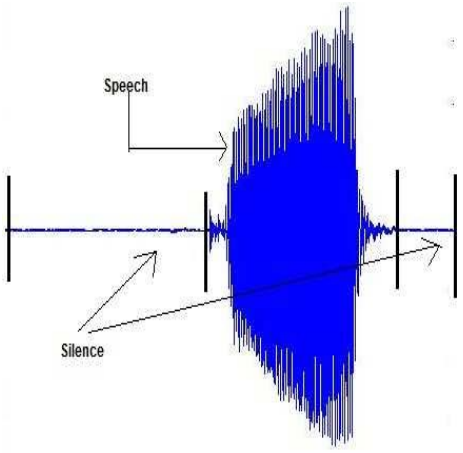

Fig. 7. Signal represent Speech and Silence segment

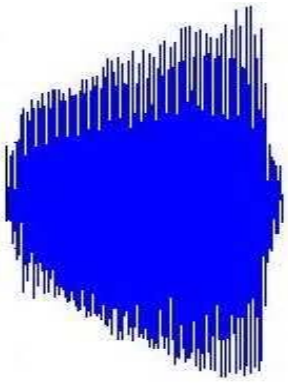

Fig. 8. Separated Speech Segment

The results generated are the average values for 30 training phonemes applied to an RNN with 35 hidden neurons.

B. Results of Speech segmentation by conventional RNN

ANN is a non- parametric prediction tool that can be used for a host of pattern classification/application including speech recognition [9] [10] [11]. One of the most commonly used ANN structures for pattern classification is the Multi-Layer Perceptron (MLP) which have found application is a host of work related to speech synthesis and recognition [9] [10] [11]. The MLPs work very well as an effective classifier for vowel sounds with stationary spectra, while their phoneme discriminating power deteriorates for consonants characterized by variations of short-time spectra [12]. Feed forward MLPs are unable to deal with time varying information as seen in the speech spectra. The RNN has the ability to deal with time varying nature of the inputs for which these are found to be suitable for application like speech recognition [13]. The input layer size is equal to the length of the feature vector and the output layer is equal to the number of classes. The block diagram of RNN based speech recognizer is shown in the Figure 9.

TABLE I

Performance of RNN v/s predictor Size

Sl No	Predictor size	Normalized training time	Precision in %
1	5	0.26	74
2	10	0.31	82
3	15	0.40	84
4	20	0.47	93.6
5	25	0.50	94
6	30	0.54	93
7	35	0.67	95
8	40	1	94

TABLE II

Performance of RNN with number of hidden neurons

Sl No	Hidden layer size	Normalized training time	Precision in %
1	20	0.18	80
2	25	0.22	82
3	30	0.26	85
4	35	0.38	94
5	40	0.44	91
6	45	0.79	93.4
7	50	0.90	93
8	55	1	93.8

The Table II shows its performance with a number of hidden neurons when trained up to 1000 epochs. The RNN training is carried out using (error) back propagation with momentum (BPM) and Levenberg-Marquardth back propagation algorithms.

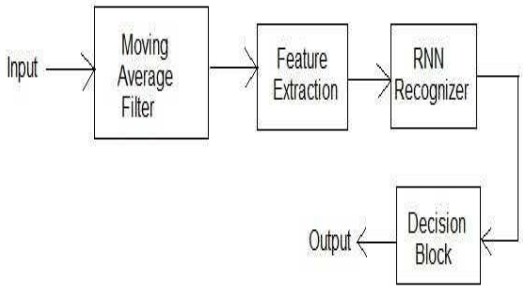

Fig. 9. RNN based Speech Recognizer

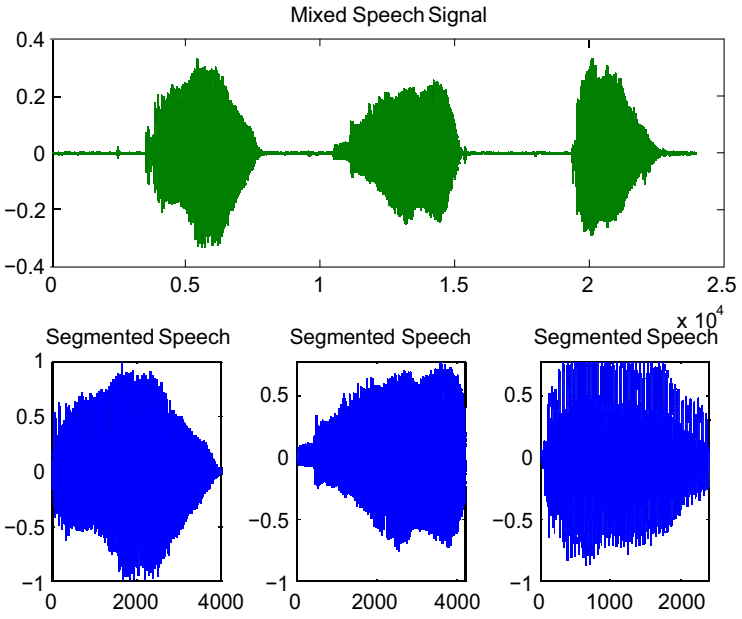

Fig. 10. Segmentation of mixed speech signal

The precision generated by both training methods are comparable but the time taken by the first one is more. But memory requirement of the second is higher. Hence, the results derived are the average values of both the methods. But for testing the RNN trained with BPM is adopted. RNN configuration with 35 hidden neurons gives the best success-rate. It requires the least amount of time to attain a success rate of around 94% within 1000 epochs. Hence, this RNN configuration is used for performing the speech recognition.

E. Results of Dynamic segmentation of speech by RNN (proposed method)

In the proposed method, conventional segmentation is replaced by dynamic segmentation approach. For proper segmentation of input speech, an optimal speech recognizer is needed. This requirement is fulfilled by the RNN based recognizer because it attain a recognition rate of above 90% [3] [8]. The results of recognition success scores and testing computational time for conventional and proposed methods are given in Table III. It may be noted from this table that the percentage recognition success score of the proposed method is 100% while that of the conventional method is 84%. Thus the proposed method may be regarded as superior to the conventional method in regard to the percentage recognition success score. The testing time of the proposed method is 0.3 seconds more than that of the conventional method due to dynamic feedback loop computation. This increase in testing time may not matter much since speed of the computation may be improved by parallel processing techniques.

TABLE III
Comparative normalized testing time performance

Item	Time	Precision
Conventional approach	0.7	.84
Proposed method	1	1

V. Conclusion and Further Direction

Here, we have proposed a dynamic segmentation method of speech samples for use with STCS for Assamese language. The dynamic segmentation method designed using RNN shows better precision than the conventional approach and also has no requirement of prior information. This way, the proposed method is more suitable than conventional approach. The "proposed" method in an expanded form can be adopted as part of a STCS in Assamese language which shall have wide spread application.

References

[1] T. Nagarajan, H. A. Murthy, R. M. Hegde: "Segmentation of speech into syllable-like units", in Proceedings of 8^{th} *European Conference on Speech Communication and Technology*, pp 2893-
2896, Geneva, 2003.

[2] M. A. Weiye: "Connectionist Vector Quantization in Automatic Speech Recognition", Doctor of Philoshopy thesis submitted to Katholieke Universiteit Leuven, Jan. 1999.

[3] R. Rabiner, B. Juang,`` Fundamentals of speech recognition", Prentice Hall, 1993.

[4] K. K. Sarma, K. Dutta and M. Sarma: "Speech Corpus Of Assamese Numerals ", 1^{st} ed., Lambert Academic Publishing, 2011.

[5] "tdil.mit.gov.in / assamesecodechartoct02.pdf" , courtesy: Prof. Gautam Baruah, Dept. of CSE, IIT Guwahati

[6] B. Yegnanarayana: "Artificial Neural Networks", 1^{st} ed., PHI, New Delhi, 2003.

[7] "http://en.wikipedia.org/wiki/Moving-average",

[8] M. Sarma , K. Dutta, K. K. Sarma: "LPC-Cepstrum Corpus of Assamese Numerals for Speech Recognition Using Recurrent Neural Network", in Proceedings of 2^{nd} IEEE *International Con-*

ference on Advances in Communication, Network, and Computing, pp 140-142, Calicut, India, 2010.

[9] D. Jurafsky, J. H. Martin: "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", 1^{st} ed., Prentice Hall, 2000

[10] A. K. Paul, D. Das, and Md. M. Kamal: "Bangla Speech Recognition System Using LPC and ANN", Proccedings of 7^{th} *International Conference on Advances in Pattern Recognition*, pp 171-174, Kolkata, Feb. 04-06, 2009.

[11] G. Dede and M. H. Sazl: "Speech recognition with artificial neural networks", *Digital Signal Processing*, vol. 20, issue 3, pp 763-768, May 2010

[12] A. M. Ahmad, S. Ismail, D. F. Samaon: "Recurrent Neural Network with Backpropagation through Time for Speech Recognition", in Proccedings of *Intemational Symposium on Communications and Information Technologies* (ISCIT), pp 98 - 102, Sapporo, Japan, Oct. 2004

[13] S. Haykin: "Neural Networks A Comprehensive Foundation", 2^{nd} ed., Pearson Education, New Delhi, 2003.